Centre for Population Change Working Paper

Number 40

Grouped time-series forecasting with an application to regional infant mortality counts

Han Lin Shang

Peter W. F. Smith

November 2013

ISSN2042-4116

ABSTRACT

We describe two methods for forecasting a grouped time series, which provides point forecasts that are aggregated appropriately across different levels of the hierarchy. Using the regional infant mortality counts in Australia, we investigate the one-step-ahead to ten-step-ahead point forecast accuracy, and examine statistical significance of the point forecast accuracy between methods. Furthermore, we introduce a novel bootstrap methodology for constructing point-wise prediction interval in a grouped time series, investigate the interval forecast accuracy, and examine the statistical significance of the interval forecast accuracy.

KEYWORDS

bottom-up forecasts; hierarchical forecasting; optimal combination forecasts; reconciling forecasts

EDITORIAL NOTE

Dr Han Lin Shang is a Research Fellow at the ESRC Centre for Population Change (CPC), working with his colleagues and co-authors of this paper on developing a dynamic population model for the UK.

Professor Peter Smith is Professor of Social Statistics and leads the CPC work package 'modelling population growth and enhancing the evidence base for policy'. Peter's research interests include graphical modelling, exact inference and models for longitudinal data.

Corresponding author: Han Lin Shang, H.Shang@southampton.ac.uk

The ESRC Centre for Population Change Working Paper Series is edited by Teresa McGowan.

ESRC Centre for Population Change

The ESRC Centre for Population Change (CPC) is a joint initiative between the Universities of Southampton, St Andrews, Edinburgh, Stirling, Strathclyde, in partnership with the Office for National Statistics (ONS) and the National Records of Scotland (NRS). The Centre is funded by the Economic and Social Research Council (ESRC) grant number RES-625-28-0001.

Website | Email | Twitter | Facebook | Mendeley

GROUPED TIME-SERIES FORECASTING WITH AN APPLICATION TO REGIONAL INFANT MORTALITY COUNTS

TABLE OF CONTENTS

1 INTRODUCTION

Advances in data collection and storage have facilitated the presence of multiple time series that are hierarchical in structure and have clusters which may be correlated. In many applications, such multiple time series can be disaggregated into many related time series which are hierarchical in structure, based on dimensions, such as gender, geography or product type. This has led to the problem of how to model and forecast such hierarchical time series.

In the field of forecasting, analyzing hierarchical or grouped time series has received increasing attention (Dunn et al. 1971, 1976, Fliedner 2001, Marcellino et al. 2003, Athanasopoulos et al. 2009, to name only a few). In macroeconomic, Stone et al. (1942) and Weale (1988) disaggregate the national economic account into production, income and outlay, and capital transactions. Production is further classified into production in Britain and production in the rest of world; income and outlay and capital transactions are each further classified into persons, companies, public corporations, general government, and rest of world. This is an example of a hierarchical time series, in which the order of disaggregation is unique. In demographic forecasting, the infant mortality counts in Australia can be disaggregated by gender; within each gender, mortality counts can be further disaggregated by geography, e.g., state. This second example is called a grouped time series, which can be thought of as hierarchical time series without a unique hierarchical structure. In other words, the infant mortality counts in Australia can also be first disaggregated by states and then by genders, thus the order is not important. In this paper, we demonstrate the construction of prediction interval with a grouped time series.

In current statistical literature, existing approaches to grouped time-series forecasting usually consider either a bottom-up method or an optimal combination method. The bottom-up method involves forecasting each of the disaggregated series at the lowest level of the hierarchy, and then using aggregation to obtain forecasts at higher levels (Kahn 1998). By using ordinary least squares estimator in the linear regression model, Hyndman et al. (2011) introduced a statistical method for optimally combining hierarchical forecasts. These two methods allow the forecasts at the bottom level to be summed consistently to the top level, without any ad-hoc adjustment. Therefore, they can potentially improve forecast accuracy, in comparison to the independent forecasts without any adjustment (e.g., Fair & Shiller 1990, Zellner & Tobias 2000, Marcellino et al. 2003, Hubrich 2005).

Despite the usefulness of grouped time-series forecasting methods, to the best of our knowledge, there is no or little work on the construction of a prediction interval for a grouped time series. However, as pointed out by Chatfield (1993, 2000), it is important to provide interval forecasts as well as point forecasts, so as to assess future uncertainty levels; enable different strategies to be planned for the range of possible outcomes; compare forecasts from different methods more thoroughly; and explore different scenarios based on different assumptions. This motivates us to propose a parametric bootstrap technique for constructing a point-wise prediction interval in a grouped time series.

The paper is organized as follows. In Section 2, we briefly review the bottom-up and optimal combination methods that are appropriate for a grouped time series. In Section 2.4, we present a parametric bootstrap method for constructing point-wise prediction intervals. Illustrated by the Australian infant mortality counts described in Section 3, we investigate the one-step-ahead to ten-step-ahead point forecast accuracy. We reveal the most accurate method in Section 4, and carry out a hypothesis testing procedure to examine if the differences in point forecast accuracy between methods are statistically significant. In Section 5, we investigate one-step-ahead to ten-step-ahead interval forecast accuracy. Also, we carried out a likelihood-ratio test to examine if the empirical coverage probability differs significantly from the nominal coverage probability at each horizon. Conclusions are given in Section 6, along with some thoughts on how the methods developed here might be further extended.

2 SOME HIERARCHICAL FORECASTING METHODS

2.1 NOTATION

Let us consider a multi-level hierarchy, where top level (or level 0) has the completely aggregated series, level 1 is the first level of disaggregation. For instance, A denotes series A at level 1; AB denotes series B at level 2 within series A at level 1, and so on. See Figure 1 for a graphical display of the two-level hierarchical structure.

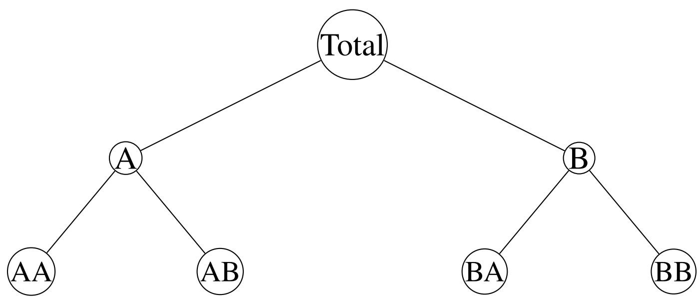


Figure 1. A two level hierarchical tree diagram.

Denote Y_t as the aggregate of all series at time $t = 1, 2, \ldots, n$. As shown in Figure 1, we have

$$Y_{\text{Total},t} = Y_{\text{A},t} + Y_{\text{B},t}, \qquad Y_{\text{A},t} = Y_{\text{AA},t} + Y_{\text{AB},t}.$$

Therefore, observations at higher levels can be obtained by summing the series below. Alternatively, we can also express the hierarchy using a matrix notation. Let $\boldsymbol{Y}_t = \left[Y_t, \boldsymbol{Y}'_{1,t}, \ldots, \boldsymbol{Y}'_{K,t}\right]'$, where $\boldsymbol{Y}_{s,t}$ represents the vector of all observations at level s at time t, and $'$ denotes the matrix transpose. Note that

$$\boldsymbol{Y}_t = \boldsymbol{S}\boldsymbol{Y}_{K,t},$$

where $\boldsymbol{S}$ is a "summing" matrix of order $m \times m_K$, and m represents the total number of series ($2^2 + 2^1 + 2^0 = 7$ for the symmetric hierarchy in Figure 1) and m_K represents the total number of bottom-level series ($2^2 = 4$). The summing matrix $\boldsymbol{S}$, which delineates how the bottom-level series are aggregated, is consistent with the hierarchical structure. For the hierarchy in Figure 1, we have

$$\begin{bmatrix} Y_t \\ Y_{A,t} \\ Y_{B,t} \\ Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{\mathbf{S}} \begin{bmatrix} Y_{AA,t} \\ Y_{AB,t} \\ Y_{BA,t} \\ Y_{BB,t} \end{bmatrix}.$$

Based on the information available up to and including time n, we are interested in computing forecasts for each series at each level, giving m independent (or base) forecasts for the forecasting period $n+h,\dots,n+w$, where h represents the forecast horizon and $w \geq h$ represents the last year of the forecasting period. We denote $\widehat{Y}_{0,h}$ as the h-step-ahead base forecast of the total series in the forecasting period, $\widehat{Y}_{\mathrm{A},h}$ as the h-step-ahead forecast of the series A, $\widehat{Y}_{\mathrm{AA},h}$ as the h-step-ahead forecast of the series AA generated for each series in the hierarchy using a suitable forecasting method, such as exponential smoothing (Hyndman et al. 2008). These base forecasts are then combined in such ways to produce final forecasts for the whole hierarchy that aggregate in a manner which is consistent with the structure of the hierarchy. We refer to these as revised forecasts and denote them as $\overline{Y}_{0,h}$ and $\overline{\boldsymbol{Y}}_{s,h}$ for level $s = 1,\dots,K$.

There are a number of ways of combining the base forecasts in order to obtain revised forecasts. The following subsections discuss two possibilities.

2.2 BOTTOM-UP METHOD

One of the commonly used methods to hierarchical forecasting is the bottom-up method (Kinney 1971, Dangerfield & Morris 1992, Zellner & Tobias 2000). This method involves first generating independent forecasts for each series at the bottom level of the hierarchy and then aggregating these upwards to produce revised forecasts for the whole hierarchy. As an example, let us consider the hierarchy of Figure 1. We first generate h-step-ahead independent forecasts for the bottom-level series, namely $\widehat{Y}_{\mathrm{AA},h}, \widehat{Y}_{\mathrm{AB},h}, \widehat{Y}_{\mathrm{BA},h}, \widehat{Y}_{\mathrm{BB},h}$. Aggregating these up the hierarchy, we get h-step-ahead forecasts for the rest of series:

$$\overline{Y}_{\mathrm{A},h} = \overline{Y}_{\mathrm{AA},h} + \overline{Y}_{\mathrm{AB},h},$$

$$\overline{Y}_{\mathrm{B},h} = \overline{Y}_{\mathrm{BA},h} + \overline{Y}_{\mathrm{BB},h},$$

$$\overline{Y}_{h} = \overline{Y}_{\mathrm{A},h} + \overline{Y}_{\mathrm{B},h}.$$

The revised forecasts for the bottom-level series are the same as the base forecasts in the bottom-up method (i.e., $\overline{Y}_{\mathrm{AA},h} = \widehat{Y}_{\mathrm{AA},h}$).

The bottom-up method can also be expressed by the summing matrix and we write

$$\overline{\boldsymbol{Y}}_h = \boldsymbol{S}\widehat{\boldsymbol{Y}}_{K,h},$$

where $\overline{\boldsymbol{Y}}_h = [\overline{Y}_{0,h}, \overline{\boldsymbol{Y}}'_{1,h}, \ldots, \overline{\boldsymbol{Y}}'_{K,h}]'$ represents the revised forecasts for the whole hierarchy, and $\widehat{\boldsymbol{Y}}_{K,h}$ represents the bottom-level forecasts.

The main advantage of this bottom-up method is that no information is lost due to aggregation. On the other hand, it may lead to inaccurate forecasts of the top-level series, when there are missing or noisy data at the bottom level (see for example, Shlifer & Wolff 1979, Schwarzkopf et al. 1988).

2.3 OPTIMAL COMBINATION METHOD

This method involves first generating base forecasts for each series. As these base forecasts are independently generated, they will not be 'aggregate consistent' (i.e., they will not sum to the group structure). The optimal combination method optimally combines the base forecasts through linear regression and generates a set of revised forecasts that are as close as possible to the independent forecasts but also aggregate consistently within the group. The essential idea is derived from the representation of h-step-ahead base forecasts for the whole of the hierarchy by linear regression. In general,

$$\widehat{\boldsymbol{Y}}_h = \boldsymbol{S}\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_h,$$

where $\widehat{\boldsymbol{Y}}_h$ is a vector of the h-step-ahead base forecasts for the whole hierarchy, stacked in the same order as for $\boldsymbol{Y}_t$; $\boldsymbol{\beta}_h = \mathrm{E}\left[\boldsymbol{Y}_{K,n+h} \middle| \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n\right]$ is the unknown mean of the base forecasts of the bottom level K, $\boldsymbol{\varepsilon}_h$ has zero mean and unknown covariance matrix $\boldsymbol{\Sigma}_h$. Note that $\boldsymbol{\varepsilon}_h$ represents the error in the above regression and should not be confused with the h-step-ahead forecast error.

Provided the base forecasts approximately satisfy the group aggregation structure (which should occur for any reasonable set of forecasts), the errors approximately satisfy the same aggregation structure as the data. That is, $\boldsymbol{\varepsilon}_h \approx \boldsymbol{S}\boldsymbol{\varepsilon}_{K,h}$, where $\boldsymbol{\varepsilon}_{K,h}$ represents the forecast errors in the bottom level. Under this assumption, Hyndman et al. (2011) show that the best linear unbiased estimator for $\boldsymbol{\beta}_h$ is

$$\widehat{\boldsymbol{\beta}}_h = \left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1} \boldsymbol{S}'\widehat{\boldsymbol{Y}}_h. \tag{1}$$

For the derivation of (1), consult Hyndman et al. (2011, Theorem 1). The revised forecasts are then given by

$$\overline{\boldsymbol{Y}}_h = \boldsymbol{S}\left(\boldsymbol{S}'\boldsymbol{S}\right)^{-1} \boldsymbol{S}'\widehat{\boldsymbol{Y}}_h,$$

which does not depend on $\boldsymbol{\Sigma}_h$.

The main advantage of this method is that it provides unbiased revised forecasts. However, its main weakness is the assumption that the errors approximately satisfy the same aggregation structure as the data.

2.4 INTERVAL FORECAST

Bootstrap techniques become important tools for assessing the parameter uncertainty, since the seminal work by Efron (1979). The monographs by Hall (1992) and Efron & Tibshirani (1993) provide detailed overview on the topic of bootstrap, while Kreiss & Paparoditis (2011) present a recent survey. Using a bootstrap technique, we take up the call of Hyndman et al. (2011) by constructing the point-wise prediction interval for a grouped time series. Our proposed method fits within the framework of parametric bootstrapping, where we produce B samples of forecasts from the fitted exponential smoothing model for each series in the bottom level. By the principle of parametric bootstrapping, the bootstrap samples are capable of mimicking the correlation within each series. Based on these B bootstrap forecasts, we assess the variability of point forecasts by constructing the prediction intervals using quantiles. As pointed out by Davidson & MacKinnon (2000), $B = 399$ would seem to be the minimum number of bootstrap replications for a test at the 0.05 level of significance.

Computationally, the `simulate.ets` function in the *forecast* package (Hyndman 2013) was utilized for simulating a number of replications of the bottom-level forecasts, from which we can then construct the series at the upper levels. By no means is exponential smoothing the only univariate time-series forecasting method, but its simplicity often provides a good forecast accuracy (see for example, Makridakis & Hibon 2000). As also noticed by Schwarzkopf et al. (1988), the independent forecasts obtained from exponential smoothing are fairly robust in correcting for possible outliers.

3 DATA SET

We apply the two grouped time-series forecasting methods to Australian infant mortality counts across different genders and states. For each series, we have yearly observations on the number of infant mortality counts from 1933 to 2003. This data set was obtained from the Australian Social Science Data Archive (`http://www.assda.edu.au/`), and is also publicly available in the *hts* package (Hyndman et al. 2013) in the R language (R Core Team 2013). Based on these observations, we are interested in forecasting regional infant mortality counts from 2004 to 2013.

The structure of the hierarchy is displayed in Table 1. At the top level, we have aggregated total infant mortality counts for the whole of Australia. At level 1, we can split this total count by gender, although we note the possibility of splitting the total counts by regions. In the third level, the total counts are disaggregated by the states and territories of Australia: New South Wales (NSW), Victoria (VIC), Queensland (QLD), South Australia (SA), Western Australia (WA), Northern Territory (NT), Australian Capital Territory and Overseas Territory (ACTOT), and Tasmania (TAS). In the bottom level, the total counts are disaggregated by the states and territories of Australia for each gender. This gives 16 series at the bottom level, and 27 series in total.

Level	Number of series
Australia	1
Gender	2
State	8
Gender $\times$ State	16
Total	27

Table 1. Hierarchy of Australian infant mortality counts.

Figure 2 shows the forecasts of regional mortality counts, using the bottom-up method. The forecasts indicate a continuing decline in infant mortality counts, due largely to improved health services. Moreover, the male infant mortality counts are higher than the female infant mortality counts. This confirms the same finding as Drevenstedt et al. (2008).

4 RESULTS OF POINT FORECAST

For each series given in Table 1, we select an optimal exponential smoothing model based on Akaike's (1974) Information Criterion. We then re-estimate the parameters of the model using a so-called rolling window approach, beginning with the model fitted using the first 61 observations (from 1933 to 1993). Forecasts from the fitted model are produced for up to ten steps ahead. We iterate this process, by increasing the sample size by one year until the end of data period in 2003. This process produces 10 one-step-ahead forecasts, 9 two-step-ahead

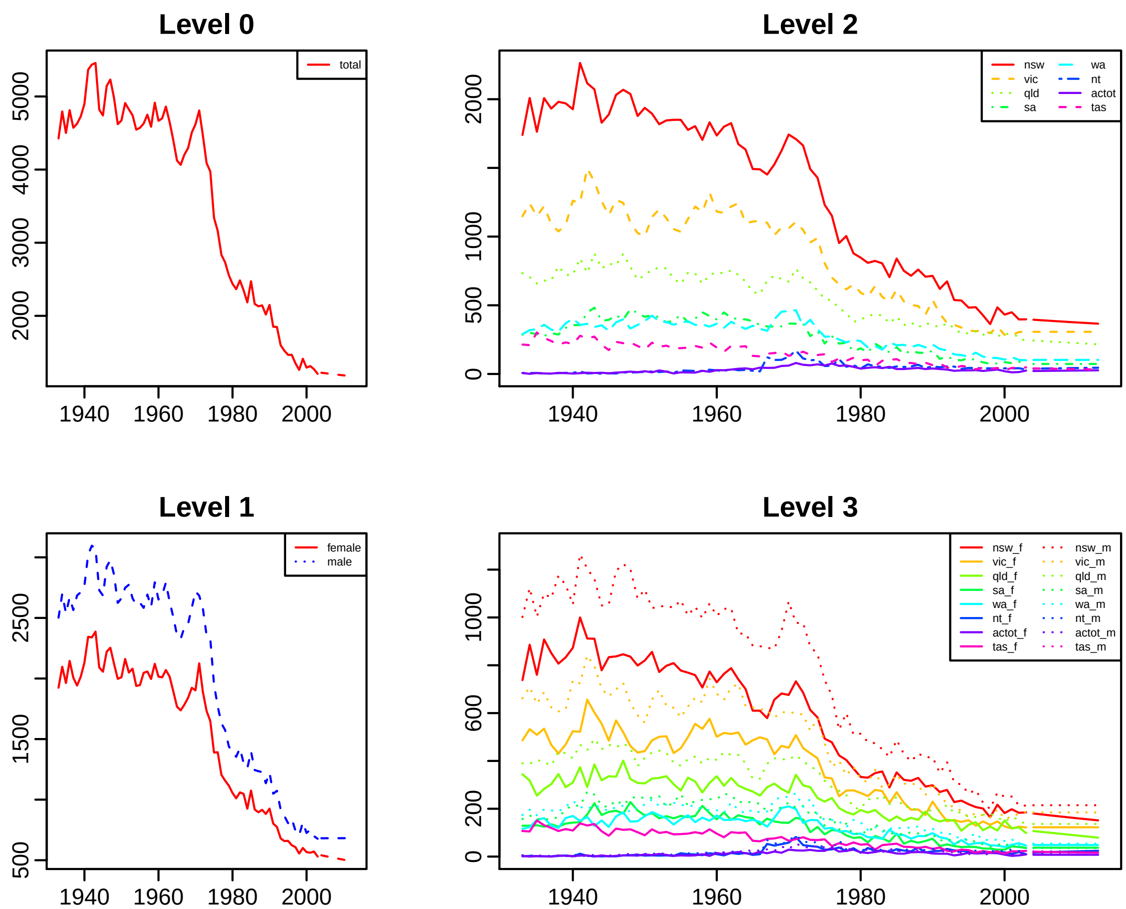


Figure 2. Point forecasts of a grouped time series using the bottom-up method. Based on the historical data from 1933 to 2003, we forecast the infant mortality counts across genders and states from 2004 to 2013.

forecasts, . . . , and 1 ten-step-ahead forecast. We use these to evaluate the out-of-sample point forecast accuracy.

In order to compare point forecast accuracy between the two grouped time-series forecasting methods, we use the mean absolute percentage error (MAPE). The MAPE is the average of percentage error, across years in the forecasting period. For each series j, it can be defined as

$$\text{MAPE}_{j,h} = \frac{1}{(11-h)} \sum_{t=n}^{n+(10-h)} \left| \frac{Y_{t+h,j} - \widehat{Y}_{t+h,j}}{Y_{t+h,j}} \right| \times 100, \qquad h = 1, \dots, 10.$$

By averaging $\text{MAPE}_{j,h}$ across 27 series, we obtain an overall assessment of the point forecast accuracy.

Table 2 contains the MAPE$_h$ for each level using the two grouped time-series methods, and an independent (base) forecasting method without reconciling forecasts. The bold entries highlight the method that performs best for the corresponding level and forecast horizon, based on the smallest MAPE. The last column contains the average MAPE across all the forecast horizons.

	Forecast horizon (h)										
	1	2	3	4	5	6	7	8	9	10	Mean
Top level: Australia (1 series)											
Base	**4.12**	**4.89**	5.06	**7.21**	9.40	7.97	6.95	9.08	8.43	6.66	6.98
Bottom-up	4.64	5.09	5.37	7.79	6.83	4.25	4.98	4.68	8.23	16.89	6.88
Combination	4.50	5.09	**4.81**	7.33	**6.16**	**3.22**	**2.36**	**0.53**	**2.00**	**1.29**	**3.73**
Level 1: Gender (2 series)											
Base	5.62	6.50	7.57	9.57	10.95	12.60	15.65	18.04	18.39	15.85	12.07
Bottom-up	5.24	**5.13**	**5.73**	7.63	7.59	**5.08**	5.73	**5.81**	8.01	16.31	7.23
Combination	**4.84**	5.20	6.07	**7.44**	**7.01**	5.94	**3.77**	5.94	**5.96**	**12.33**	**6.45**
Level 2 : State (8 series)											
Base	**13.54**	15.50	17.27	19.45	22.30	24.20	28.10	29.98	31.09	**37.96**	23.94
Bottom-up	13.81	**15.09**	16.94	19.33	22.36	24.17	28.63	31.94	33.04	40.55	24.59
Combination	14.16	15.26	**15.90**	**18.21**	**20.87**	**23.55**	**23.60**	**27.30**	**28.85**	40.47	**22.82**
Bottom level: Gender × State (16 series)											
Base	19.08	21.74	24.77	27.45	30.42	**32.66**	36.43	40.09	**37.00**	**44.50**	**31.41**
Bottom-up	19.08	21.74	24.77	27.45	30.42	**32.66**	36.43	40.09	**37.00**	**44.50**	**31.41**
Combination	**18.95**	**20.32**	**22.38**	**23.37**	**28.25**	32.76	**32.89**	**38.31**	42.10	54.96	31.43
Average across all levels : (27 series)											
Base	**10.59**	12.16	13.67	15.92	18.27	19.36	21.78	24.30	23.73	**26.24**	18.60
Bottom-up	10.69	11.76	13.20	15.55	16.80	16.54	18.94	20.63	21.57	29.56	17.53
Combination	10.61	**11.47**	**12.29**	**14.09**	**15.57**	**16.37**	**15.66**	**18.02**	**19.73**	27.26	**16.11**

Table 2. MAPE for out-of-sample forecasts of the grouped time-series methods applied to Australian infant mortality counts. The bold entries highlight the method that performs best for the corresponding level and forecast horizon.

From this empirical study, we find that the optimal combination method and bottom-up method outperform the independent forecasts at the top two levels. In the top level, the optimal combination method performs the best for long-term forecasts, whereas the independent method generally performs the best for short-term forecast horizons. In the bottom level, the independent forecasts and bottom-up forecasts are the same. Based on the overall MAPE across all hierarchical levels, the good performance of the bottom-up and optimal combination methods can be attributed to the fact that the data have strong trends, even at the bottom level. With more noisy data, it may not be easy to extract the signal at the bottom level, and this may lead to inferior performance of the grouped time-series methods.

To formally test whether the differences in forecast accuracy among the alternative methods are significant, we perform the Friedman's (1937) test. Friedman's test is a nonparametric analog of variance for a randomized block design, which can be considered as nonparametric version of a one-way ANOVA with repeated measures. The errors are ranked from the smallest to the largest across the series and the sum of the ranks for each method is compared. When two errors are the same, a mean rank is then assigned. These ranks, denoted by R_j, are given in Table 3. If the forecast accuracy is different among methods, there would be a significant difference in the sum of the ranks of at least one method.

The Friedman's test statistic is given by

$$\frac{12}{HK(K+1)}\sum R_j^2 - 3H(K+1),$$

where $K = 3$ is the number of methods considered and $H = 10$ is the number of horizons. Friedman's test is a procedure based on within-block ranks, and the test statistics has approximately a χ^2_{K-1} distribution with $K-1$ degrees of freedom, when the null hypothesis is true (see Demšar 2006, for details). By comparing the Friedman test statistics with the critical values, there is a significant difference among methods in the top two levels; see Table 3.

Having identified the statistical significance among methods, we perform the Nemenyi's (1963) test, which is a post-hoc pairwise test intended to determine which method is significantly different from the others. The Nemenyi test is two-sided with the null hypothesis being two methods give similar point forecast accuracy. The forecast accuracy of two methods is

	Forecast horizon (h)													
	1	2	3	4	5	6	7	8	9	10	R_j	R_j^2	$\sum R_j^2$	F
Top level: Australia (1 series)														
Base	1	1	2	1	3	3	3	3	3	2	22	484		
Bottom-up	3	2.5	3	3	2	2	2	2	2	3	24.5	600.25		
Combination	2	2.5	1	2	1	1	1	1	1	1	13.5	182.25		
Test statistics													1266.5	6.65
Level 1: Gender (2 series)														
Base	3	3	3	3	3	3	3	3	3	2	29	841		
Bottom-up	2	1	1	2	2	1	2	1	2	3	17	289		
Combination	1	2	2	1	1	2	1	2	1	1	14	196		
Test statistics													1326	12.6
Level 2 : State (8 series)														
Base	1	3	3	3	2	3	2	2	2	1	22	484		
Bottom-up	2	1	2	2	3	2	3	3	3	3	24	576		
Combination	3	2	1	1	1	1	1	1	1	2	14	196		
Test statistics													1256	5.6
Bottom level: Gender × State (16 series)														
Base	2.5	2.5	2.5	2.5	2.5	1.5	2.5	2.5	1.5	1.5	22	484		
Bottom-up	2.5	2.5	2.5	2.5	2.5	1.5	2.5	2.5	1.5	1.5	22	484		
Combination	1	1	1	1	1	3	1	1	3	3	16	256		
Test statistics													1224	2.4

Table 3. Ranks of forecast errors by method and horizon. The critical value is $\chi^2_{0.95,2} = 5.99$.

significantly different if the corresponding average ranks differ by at least the critical difference

$$q_\alpha \sqrt{\frac{K(K+1)}{6H}},$$

where critical values q_α are given in (Demšar 2006, Table 5). Based on the test statistics, we can calculate its corresponding p-value shown in Table 4. We conclude that

(1) at the top level, the bottom-up method differs significantly from the optimal combination method;

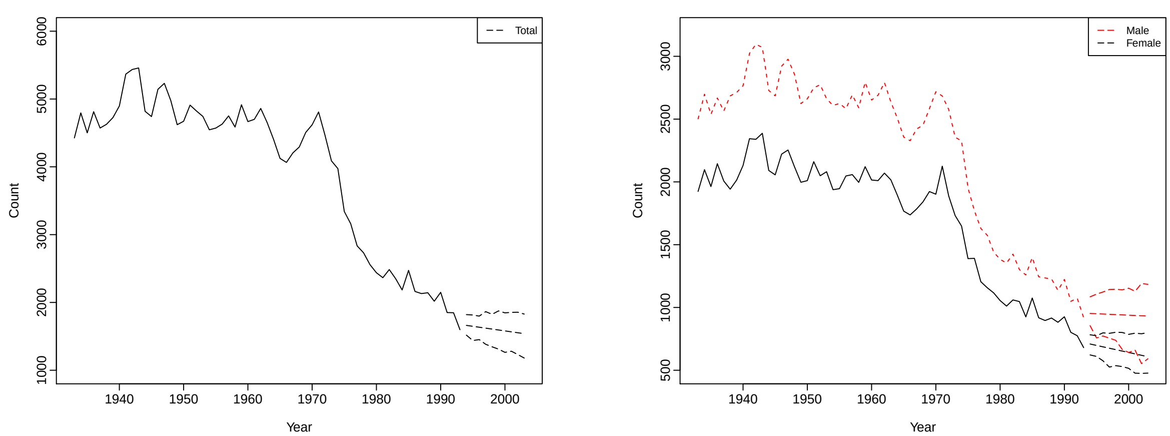
(2) at the first level, the bottom-up and optimal combination methods differ significantly from the independent forecasting method.

	Base	Optimal combination
Top level: Australia		
Optimal combination	0.1316	—
Bottom-up	0.8381	**0.0341**
Level 1: Gender		
Optimal combination	**0.0023**	—
Bottom-up	**0.0199**	0.7805

Table 4. p-values of the Nemenyi's test statistics to test statistical significance of the point forecast accuracy among methods.

5 RESULTS OF INTERVAL FORECAST

As described in Section 2.4, we construct point-wise prediction interval using a parametric bootstrap approach. This approach produces a number of bootstrapped point forecasts from the fitted exponential smoothing model for each series in the bottom level. Based on these B bootstrap forecasts, we assess the variability of point forecasts by constructing the prediction intervals using quantiles. For instance, Figure 3 displays the 80% point-wise prediction interval of the regional infant mortality counts from 2004 to 2013 at the top two levels.



(a) Level 0 series

(b) Level 1 series

Figure 3. 80% point-wise prediction intervals of a hierarchical time series.

Given a sample path $\{\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n\}$ where $\boldsymbol{Y}_t$ is a column vector of series across the entire hierarchy, we constructed the h-step-ahead interval forecasts, with the lower bound denoted by

$\boldsymbol{L}_{n+h|n}(p)$ and upper bound denoted by $\boldsymbol{U}_{n+h|n}(p)$, where p is the nominal coverage probability. Given the holdout time series in the forecasting period and the point-wise interval forecasts, the indicator variable is defined as

$$I_{n+h,j} = \begin{cases} 1 & \text{if } Y_{n+h,j} \in [L_{n+h|n,j}(p), U_{n+h|n,j}(p)]; \\ 0 & \text{if } Y_{n+h,j} \notin [L_{n+h|n,j}(p), U_{n+h|n,j}(p)], \qquad j = 1, \dots, m, \end{cases}$$

where $m = 27$ represents the total number of series in a hierarchy. Let $\boldsymbol{I}_{n+h}$ be a column vector of indicator variables for h-step-ahead forecast, and $\boldsymbol{Y}_{n+h}$ be a column vector of series across the entire hierarchy. In our data set, $\boldsymbol{Y}_{n+h} = [Y_{n+h,\text{total}}, Y_{n+h,\text{female}}, Y_{n+h,\text{male}}, Y_{n+h,\text{female}_{\text{nsw}}}, \dots, Y_{n+h,\text{female}_{\text{tas}}}, Y_{n+h,\text{male}_{\text{nsw}}}, \dots, Y_{n+}$

Recall that the training sample period is from 1933 to 1993 (observations 1 to n), while the testing sample period is from 1994 to 2003 (observations $n + 1, \dots, n + 10$). For each forecast horizon, the empirical coverage probability is defined as

$$1 - \frac{\sum_{l=n}^{n+(10-h)} \sum_{j=1}^{m} I_{l+h,j}}{m \times (11 - h)}, \qquad h = 1, \dots, 10.$$

In Table 5, we present the empirical coverage probabilities for the one-step-ahead to ten-step-ahead interval forecasts, using the bottom-up method. From $h = 1$ to $h = 10$, we find that the empirical coverage probabilities are reasonably close to the nominal coverage probability of 0.8.

Forecast horizon	1	2	3	4	5	6	7	8	9	10
Empirical coverage	0.71	0.72	0.75	0.69	0.64	0.73	0.72	0.69	0.72	0.74

Table 5. Empirical coverage probabilities for the one-step-ahead to ten-step-ahead interval forecasts using the bottom-up method, at the nominal coverage probability of 0.8.

To formally test if the empirical coverage probability differs from the nominal coverage probability significantly, we performed log likelihood-ratio (LR) test statistics (see Christoffersen 1998, for detail). Christoffersen (1998) proposed a test for unconditional coverage, a test for independence of indicator sequence, and a joint test of conditional coverage and independence. It is the joint test that we applied to our one-step-ahead to ten-step-ahead indicator sequence.

At the nominal coverage probability of 0.8, the log LR test statistics are given in Table 6. The log LR test statistics are compared with the critical value of $\chi^2_{0.95,2} = 5.99$ at the 5% level of significance. The log LR test statistics is greater than the critical value, only when $h = 5$. Therefore, overall there is strong evidence that the parametric bootstrap method has good coverage properties.

Forecast horizon	1	2	3	4	5	6	7	8	9	10
Log LR test statistics	5.73	4.55	1.87	3.24	9.23	5.28	5.94	4.03	2.55	5.01

Table 6. Log likelihood-ratio test statistics for the one-step-ahead to ten-step-ahead interval forecasts at the nominal coverage probability of 0.8. The critical value is 5.99 at the 5% level of significance.

6 CONCLUSIONS

This article begins by revisiting two methods for modeling and forecasting a grouped time series. The bottom-up method models and forecasts data series at the bottom level, and then aggregates to the top level. The optimal combination method considers the problem of group forecasting from a regression perspective, and it uses the ordinary least squares to find the optimal regression coefficients, based on which forecasts are obtained.

Illustrated by the regional infant mortality counts in Australia, we demonstrate the use of bottom-up method for producing point and interval forecasts of mortality counts from 2004 to 2013. Furthermore, we compared the one-step-ahead to ten-step-ahead point forecast accuracy of each of the grouped time-series methods, and found that the optimal combination method performs the best with the smallest MAPE, averaged across different levels. In order to examine if the differences among methods are statistically significant, we carried out the Friedman's test and Nemenyi's test across different levels of the hierarchy. At the top level, the bottom-up method differs significantly from the optimal combination method, whereas the bottom-up and optimal combination methods differ significantly from the independent forecasting method at the first level. For the lower levels, there is no significant difference among methods.

The main contribution of this paper is to propose a parametric bootstrap method for constructing the point-wise prediction intervals. For one-step-ahead to ten-step-ahead interval forecasts, our method produces the empirical coverage probabilities that are close to the nominal coverage probability. A log likelihood-ratio test is implemented to examine if the empirical coverage probability differs significantly from the nominal coverage probability. At the short and long forecast horizons, we infer that the empirical coverage probability does not differ significantly from the nominal coverage probability.

There are many ways in which the paper can be further extended, and we briefly mention a few. First, we demonstrated the group time-series forecasting methods by the infant mortality counts in Australia, but the methodology presented can easily be extended to forecast age-specific mortality counts by adding age as an additional level of the group. Second, the methodology can be applied to cause-specific mortality, considered in Murray & Lopez (1997) and Girosi & King (2008). Third, since the forecast accuracy depends on the unknown signal-to-noise ratio of the bottom-level series, the forecast accuracy may be improved by averaging the forecasts produced by different hierarchical forecasting methods. Finally, the idea of grouped time series can be extended to functional time series, where each series is a time series of functions.

Implementation of the two grouped time-series methods are straightforward using the readily available R package *hts* (Hyndman et al. 2013). This package provides the point forecasts, while the computational code for constructing point-wise prediction interval can be obtained upon request from the authors.

REFERENCES

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19 (6), 716-723.

Athanasopoulos, G., Ahmed, R.A. and Hyndman, R.J. (2009) Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting*, 25 (1), 146-166.

Chatfield, C. (1993) Calculating interval forecasts. *Journal of Business & Economic Statistics,* 11 (2), 121-135.

Chatfield, C. (2000) Time-Series Forecasting. *Chapman & Hall/CRC*, Boca Raton, Florida.

Christoffersen, P.F. (1998) Evaluating interval forecasts. *International Economic Review,* 39 (4), 841-862.

Dangerfield, B.J. and Morris, J.S. (1992) Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting,* 8 (2), 233-241.

Davidson, R. and MacKinnon, J.G. (2000) Bootstrap tests: How many bootstraps? *Econometric Reviews,* 19 (1), 55-68.

Demsar, J. (2006) Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7, 1-30.

Drevenstedt, G.L., Crimmins, E.M., Vasunilashorn, S. and Finch, C.E. (2008) The rise and fall of excess male infant mortality. *Proceedings of the National Academy of Sciences of the United States of America,* 105 (13), 5016-5021.

Dunn, D.M., William, W.H. and Spiney, W.A. (1971) Analysis and prediction of telephone demand in local geographic areas. *Bell Journal of Economics and Management Science,* 2 (2), 561-576.

Dunn, D.M., Williams, W.H. and DeChaine, T.L. (1976) Aggregate versus subaggregate models in local area forecasting. *Journal of the American Statistical Association,* 71 (353), 68-71.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Annals of Statistics,* 7 (1), 1-26.

Efron, B. and Tibshirani, R.J. (1993) An Introduction to the Bootstrap. *Chapman & Hall,* London.

Fair, R.C. and Shiller, R.J. (1990) Comparing information in forecasts from econometric models. *The American Economic Review,* 80 (3), 375-389.

Fliedner, G. (2001) Hierarchical forecasting: Issues and use guidelines. *Industrial Management and Data Systems,* 101 (1), 5-12.

Friedman, M. (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association,* 32 (200), 675-701.

Girosi, F. and King, G. (2008) Demographic Forecasting. *Princeton University Press,* Princeton.

Hall, P. (1992) The Bootstrap and Edgeworth Expansion. *Springer Series in Statistics,* Springer-Verlag, New York.

Hubrich, K. (2005) Forecasting euro area ination: Does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting*, 21 (1), 119-136.

Hyndman, R.J. (2013) forecast: Forecasting functions for time series and linear models. *R package version 4.06.* URL: http://CRAN.R-project.org/package=forecast

Hyndman, R.J., Ahmed, R.A., Athanasopoulos, G. and Shang, H.L. (2011) Optimal combination forecasts for hierarchical time series. *Computational Statistics and Data Analysis,* 55 (9), 2579-2589.

Hyndman, R.J., Ahmed, R.A. and Shang, H.L. (2013) hts: Hierarchical and grouped time series. *R package version 3.03.* URL: http://CRAN.R-project.org/package=hts

Hyndman, R., Koehler, A., Ord, J. and Snyder, R. (2008) Forecasting with Exponential Smoothing: the State-Space Approach. *Springer,* New York.

Kahn, K.B. (1998) Revisiting top-down versus bottom-up forecasting. *The Journal of Business Forecasting,* 17 (2), 14-19.

Kinney, W.R. (1971) Predicting earnings: Entity versus subentity data. *Journal of Accounting Research,* 9 (1), 127-136.

Kreiss, J.P. and Paparoditis, E. (2011) Bootstrap methods for dependent data: A review. *Journal of the Korean Statistical Society,* 40 (4), 357-378.

Makridakis, S. and Hibon, M. (2000) The M3-competition: Results, conclusions and implications. *International Journal of Forecasting,* 16 (4), 451-476.

Marcellino, M., Stock, J.H. and Watson, M.W. (2003) Macroeconomic forecasting in the Euro area: Country specific versus area-wide information. *European Economic Review,* 47 (1), 1-18.

Murray, C.J.L. and Lopez, A.D. (1997) Alternative projections of mortality and disability by cause 1990-2020: Global burden of disease study. *The Lancet,* 349 (9064), 1498-1504.

Nemenyi, P.B. (1963) Distribution-free Multiple Comparisons. *PhD thesis,* Princeton University.

R Core Team (2013) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing,* Vienna, Austria. ISBN 3-900051-07-0. URL: http://www.R-project.org/

Schwarzkopf, A.B., Tersine, R.J. and Morris, J.S. (1988) Top-down versus bottom-up forecasting strategies. *International Journal of Production Research,* 26 (11), 1833-1843.

Shlifer, E. and Wolff, R.W. (1979) Aggregation and proration in forecasting. *Management Science,* 25 (6), 594-603.

Stone, R., Champernowne, D.G. and Meade, J.E. (1942) The precision of national income estimates. *The Review of Economic Studies,* 9 (2), 111-125.

Weale, M. (1988) The reconciliation of values, volumes and prices in the national accounts. *Journal of the Royal Statistical Society, Series A*, 151 (1), 211-221.

Zellner, A. and Tobias, J. (2000) A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting,* 19 (5), 457-469.

ESRC Centre for Population Change
Building 58, Room 2001
Faculty of Social and Human Sciences
University of Southampton
SO17 1BJ

T: +44 (0)2380 592579
E: cpc@soton.ac.uk
www.cpc.ac.uk

To subscribe to the CPC newsletter and keep up-to-date with research activity, news and events, please register online: www.cpc.ac.uk/newsletter

You can also follow CPC on Twitter, Facebook and Mendeley for our latest research and updates:

www.facebook.com/CPCpopulation

www.twitter.com/CPCpopulation

www.mendeley.com/groups/3241781/centre-for-population-change

The ESRC Centre for Population Change (CPC) is a joint initiative between the University of Southampton and a consortium of Scottish universities including St Andrews, Edinburgh, Stirling and Strathclyde, in partnership with the Office for National Statistics and National Records of Scotland.